Towards Analyzing and Predicting the Experience of Live Performances with Wearable Sensing

Ekin Gedik, Laura Cabrera-Quiros, Claudio Martella, Gwenn Englebienne, and Hayley Hung

Abstract—We present an approach to interpret the response of audiences to live performances by processing mobile sensor data. We apply our method on three different datasets obtained from three live performances, where each audience member wore a single tri-axial accelerometer and proximity sensor embedded inside a smart sensor pack. Using these sensor data, we developed a novel approach to predict audience members' self-reported experience of the performances in terms of enjoyment, immersion, willingness to recommend the event to others, and change in mood. The proposed method uses an unsupervised method to identify informative intervals of the event, using the linkage of the audience members' bodily movements, and uses data from these intervals only to estimate the audience members' experience. We also analyze how the relative location of members of the audience can affect their experience and present an automatic way of recovering neighborhood information based on proximity sensors. We further show that the linkage of the audience members' bodily movements is informative of memorable moments which were later reported by the audience.

Index Terms—Human behavior, wearable sensors, proximity sensing, accelerometers, audience response, arts, dance

1 Introduction

Institutions that organize live performances increasingly require being able to quantify the response to the service they provide. Such quantification enables institutions to design more targeted events, make better monetary decisions and offer enhanced experiences to their public. Quantitative data about audience response should eventually allow us to demonstrate the contribution of artistic performances to the individual audience members' well-being. While art and cultural events may appear to be a luxury to have in a society, numerous studies have shown their benefits for stimulating the social life of public spaces [1], health and mental well-being [2], [3], [4], [5] and perceived quality of life [6]. In this paper, we investigate ways to automatically measure the audience members' response to a live performance, in real-time, as a means to enhance it, both for consumers and practitioners.

According to the appraisal theory, one's evaluation of a situation causes related affective responses [7]. In other words, a person's appraisal of an event will be reflected in the emotional responses they exhibit throughout the event itself. In this study, we present a method that uses this connection to detect an audience's appraisal of a live performance, based on the assumption that audience members' individual and joint body movements capture some form of affective response. We will be using a language similar to the one used in implicit tagging literature [8] to distinguish between self-reported evaluations of the event and immediate responses obtained through sensing. Questionnaire answers provide explicit responses by the participants and are indicative of their reappraisal of the event. We use the term reappraisal since questionnaires are filled in after the event finishes. Sensors, on the other hand, capture immediate responses and act as implicit cues for the appraisal of the event. We use the term implicit for evaluations obtained through sensing since it exploits the non-verbal reactions of the participant instead of direct responses. Thus, we aim to automatically predict the participants' explicit reappraisal of the event, from sensor recordings that capture their non-verbal reactions during the event. We do not explicitly detect any affective tags or emotional states but we try to connect immediate body movements to explicit evaluations of the event.

The automatic detection of people's affective state is a widely studied topic in affective computing, with a majority of the literature focusing on facial expressions [9] and/or speech [10]. However, these studies are typically conducted in controlled environments and have limitations when compared to real-life performances, both in terms of the data acquisition (high-quality video and audio collection) and of the generation (posed facial expressions, carefully designed stimuli). The practical characteristics of real-world performances are different from the pre-designed lab experiments and introduce important restrictions on the use of aforementioned modalities. For example, robustly detecting audience members' facial expressions in a dark concert hall from video input is a challenging task. Previous studies have shown, however, that body movements also convey affective expressions which might be exploited for the detection of emotional states [11], [12]. Even though most of the existing studies investigating affective body expressions use either video [13], [14], motion capture [15] or pressure sensors [16], we show that it is possible to capture enough of these body movements through the commonly available wearable accelerometers that are suitable for audiences in real-world settings.

Interestingly, in live performances, multiple people are simultaneously exposed to the same stimuli. This makes it possible to analyze and exploit the collective spontaneous response to the stimuli. It has been shown that the link between multiple people's responses can be exploited to detect salient moments of movies using physiological sensing [17] and, building on these findings, we propose a novel method to measure the audience's collective response to live performances. In contrast to prior work that exploits fairly reliable but less pervasive biosignals or physiological sensing [18], [19], we show that individual and collective body movement patterns of audience members, as measured through the accelerometers, can also be used to measure affective responses to a performance. The proposed method exploits the linkage between audience members' body movement to detect distinctive time intervals in the performance. Individual movement patterns of participants in these distinctive parts are then used to classify the general evaluation of the performance.

By working closely for the last 2 years with Holland Dance (HD), an organization whose role is to promote dance in The Netherlands, we have identified two key challenges to measuring an audience's response to live performances: the limits of surveys and the difficulty to obtain detailed responses on a large scale. Survey responses must be obtained after the performance, at a time when audience members are not necessarily eager to fill in questionnaires, and they do not capture the audience's spontaneous response to specific moments of the performance. Even when survey responses are available, a typical

- *E. Gedik and H. Hung are with the Department of Intelligent Systems, Technical University of Delft, Delft 2628 CD, The Netherlands. E-mail: {egedik, hhung}@tudelft.nl.*
- *L. Cabrera-Quiros is with the Department of Intelligent Systems, Technical University of Delft, Delft 2628 CD, The Netherlands, and the Instituto Tecnólogico de Costa Rica, Cartago 30101, Costa Rica. E-mail: l.c.cabreraquiros@tudelft.nl.*
- *C. Martella is with the Google, 1-13 St Giles High St, London WC2H 8AG, UK. E-mail: claudio.martella@vu.nl.*
- *G. Englebienne is with Twente University, Enschede 7522 NB, The Netherlands. E-mail: g.englebienne@utwente.nl.*

Manuscript received 4 Apr. 2018; revised 28 Aug. 2018; accepted 1 Oct. 2018. Date of publication 16 Oct. 2018; date of current version 1 Mar. 2021.
(Corresponding author: Ekin Gedik.)
Recommended for acceptance by Y.-H. Yang.
Digital Object Identifier no. 10.1109/TAFFC.2018.2875987

Likert scale cannot provide detailed insights into what aspects of a performance could have triggered someone to like or dislike it. One way to circumvent this problem involves using free text answers, which can provide richer information about someone's experience, but these need to be manually processed, they are harder to aggregate statistically, and they are subject to interpretation. Interviews are another possibility and provide a very rich medium for those few audience members who are willing to spend more time. They are, therefore, at best limited to an even smaller subset of an entire audience and do not provide an unbiased sample of the audience.

We address these challenges by making the following novel contributions in this study: we show, using two real-life events that (i) when people are watching a live performance, their spontaneous reactions can be captured with a standard accelerometer, (ii) some moments of collective reaction correspond to memorable events of high affective output in the performance as can be verified by survey responses, (iii) audience members' reactions can be used to predict their enjoyment of the performance, whether they felt immersed in the experience, would recommend it to others, or thought the performance changed their mood positively. In addition, we could not rule out that (iv) the physical distance between audience members and whether they joined the event as acquaintances might have an effect on the similarity of their evaluation of the event, but (v) found that we that we can approximately identify the side neighbours of audience members with an acceptable performance using neighbor sensing to take this into account.

2 Related Work

To view the measurement of responses from the perspective of appraisal theory, where affective responses are considered to be linked to the final evaluation [7], it is important to first consider basic automatic affect recognition. A large number of studies have been published on this topic in the last decades [10]. Most of the early work focused on video and/or audio inputs [9], used datasets of single input modalities [20], included a limited set of deliberate affective displays [21], and were recorded under highly constrained and artificially generated conditions [22]. More recent studies, on the other hand, generally aim to detect spontaneous affective displays [23], prefer to use multimodal information [24] and focus on detection of non-basic affective states [14].

New cues have started to gain importance in affect recognition; bodily expressions being one. The use of bodily expressions for affect recognition is supported by existing work in social psychology that shows the strong connection between body movements and affective expressions [25], [26]. The increasing availability of whole-body sensing technologies made it feasible to investigate the recognition of bodily expressions for affect perception and detection. This is reflected in the increasing number of studies that are discussed in recent surveys [11], [12] which rely on various approaches for capturing bodily expression such as computer vision [13], [14], motion capture [15] and pressure sensors [16], and generally aim to automatically map bodily expressions into well-known affective states. These affective states might be categorical [27] or continuous [28]). Most datasets used in such studies include acted bodily expressions [27], [28], however, focus is being shifted to real life data [29]. The methodology tends to be similar, where features are extracted from sensor data, followed by the training of statistical models for automatic affect detection. One key distinction between these and our approach is that we do not try to discriminate between types of bodily movements or map them to affective states.

Existing literature on the evaluation of events traditionally investigates the response of an audience to a live performance using self-reports, such as surveys and interviews [30], [31]. Digital technologies can overcome some limitations of surveys and interviews and give more direct and fine-grained insights into the response of an audience. For example, mobile computing and the explosion in popularity of social media such as Twitter have broadened the reach of a live performance, as fans comment and post information and opinions live to the online community [32]. Practitioners are interested in measuring the activity of their audience in social media, both to understand their response and to leverage their activities as marketing tools for their performances [33], [34]. For example, some theaters, including Broadway, have experimented with "tweet-seats" reserved for customers who promised to tweet about the performance live [35].

Rather less pervasive sensor technologies have also been used to overcome the granularity issues of surveys. For example, work in neuroaesthetics uses fMRI scanning to relate viewer responses to the aesthetics of the performance [36], [37], [38]. Other work used the tracking of eye gaze from video to distinguish novice from expert observers of dance [39]. Some work used physiological sensing such as galvanic skin response (GSR) sensors to measure the arousal of individuals watching a video of a dance performance and investigated its relationship with the individuals' self-reports [40], while others have used GSRs to measure the response to other types of live performance, such as comedy [41] and movies in a cinema [19]. One specific example we would like to point out is the work of Chenes et al., which used GSRs to detect highlights in movie scenes [17], and focused on exploiting the inter-user physiological linkage calculated with simple correlation in sliding windows over pairs of participants' GSR readings. This study shows that when people are exposed to the same stimuli (even at different times), they tend to give synchronous physiological responses which can be used to detect salient parts of those stimuli. We build our study on a similar base where we show that such linkage can also computed with body movements, yielding a similar result.

These attempts show an increasing interest in quantifying the experience of live performances, but their approaches would be hard to apply in real settings. Unlike these approaches, we advocate the use of pervasive sensors which are readily available in smart phones. As such, they enable less obtrusive measurements, on a massive scale, compared to those obtained via physiological sensing. This makes them much more readily deployable and vastly increases their practical use.

In this work, we rely on acceleration and proximity sensors to measure people's reactions to live performances. These sensors have thus far been limited to measuring very different phenomena such as the recognition of outdoor [42] or household activities [43], and the detection of medically relevant events [44], [45]. These all focus resolutely on physical activities where the behavior can be represented directly by quite specific body movements.

The above-mentioned work measures behavior in environments that are far less challenging than a theater, where the audience sits in silence and where the link between activity and behavior is not as direct. The most similar work to our own was presented by Englebienne and Hung [46] who found that they were able to identify professors and non-professors from their behavior in an inaugural lecture. Although they were sitting, the small movements made in reaction to the parts of the lecture demonstrated implicit responses of interest to particular moments during the lecture. Other closely related work was presented by Bao et al. [47] who investigated how to sense the implicit responses of users watching movies on a tablet. Using a multimodal approach, they were able to predict the user's ratings of the movies they watched. However, in this case, the user was alone and was not inhibited by the social norms usually adhered to in a public space.

Previous work using proximity sensors to study the interactions between individuals used approaches similar to complex network analysis. These sensors have been used for the analysis of social interactions in crowded settings[48], detecting different communities in an ICT conference [49] and discovering spatio-temporal relationships in the context of crowd dynamics [50]. While these studies show that

social relationships between individuals can be captured by means of spatio-temporal information, they rely on heterogeneous and dynamic inter-personal distances and orientations. None focus on the spatio-temporal relationship information in the context of live performances, which is paradoxically made very complicated by to the rigid grid structure of the seating arrangement.

3 Data Collection

3.1 Dataset 1: Dance Performance

The Sensor Set-Up. This study took place during a live dance performance that lasted almost an hour and a half. It consisted of mainly dancing, interspersed with monologues by the performers. The music was based on live cello arrangements and pre-recorded songs. We instrumented 41 participants watching the performance with triaxial accelerometers. The accelerometers were located in a custom-made device hung around each participant's neck, which recorded acceleration at 20 Hz and were kept synchronized to a global time through wireless network communication. The wireless radio module additionally broadcasted the device's unique identifier (ID), every second, with a range of 2-3 meters. The reception of such a broadcast by a nearby device is considered a proximity detection. Due to various hardware malfunctions, however, only 32 devices recorded acceleration data.

In addition, the performance was recorded using a GoPro Hero +3 to manually analyze salient moments (i.e., favorite moments that were reported by the participants). We used $\sim$79 minutes of sensor data in our experiments, starting just before the first piece, when all participants are seated, and ending when the final piece of the performance finishes.

Survey Responses. All 41 participants filled in a questionnaire after the performance. These questionnaires consisted of 12 questions on four topics (three questions per topic), measuring "enjoyment", "recommendation (to a friend)", "immersion" and "mood changes". All questions used a ten-point Likert scale, where one means "I completely disagree" and ten means "I completely agree". For measuring "enjoyment", we adapted and selected questions presented in [51]; for "immersion", we selected involvement questions from the Igroup Presence Questionnaire [52]; for "recommendation" we used items from O'Brien's questionnaire [53]. Each of these questions were carefully chosen to measure each task and slightly adapted to match our scenario. We formed the questions regarding mood by ourselves. The complete set of questions asked in this questionnaire in English are listed in the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2018.2875987.

In the rest of this paper, we refer to the participants "experience" of the performance to indicate the participants' sentiment about the performance, as measured by the questionnaires.

Of the 32 participants with valid accelerometer data, 25 reported a favorite moment of the performance. Two moments were particularly memorable: the *motorcycle sequence* and the *bolero finale*, declared as favorite by 32 and 52 percent of the participants, respectively. Note that some participants declared more than one favorite moment.

3.2 Dataset 2: A Day of Wonder

The Sensor Set-Up. As a follow up, we organized a second study in the 'A day of Wonder' festival that took place at the Delft University of Technology. This one-day festival is a combination of events regarding technology, music, food and art. We focused on one specific event that comprises two adjacent sets; namely 'Tales for the Curious Mind' and 'Enhancing Classical Music'. The first set included three presentations from various researchers and designers. The first presenter talked about a minimally-invasive surgical instrument, the second one described a smart wedding dress and the final speaker introduced a micro drone (delfly). The second set was an innovative classical concert experience which started with a solo piano performance, followed by the talk of the performer and concluded with the classical music piece Zigeunerreisen, performed by a duo of violin and piano. The whole festival was free to attend and was open to the public. Participation in the data collection was voluntary and participants were allowed to leave whenever they wanted. Some of the participants were seated while others were standing.

Participants wore our custom-made sensor pack hung around their necks, recording tri-axial acceleration and proximity information with the same setup (20 Hz and synchronized globally) as Dataset 1. A GoPro Hero +3 camera recorded the stage for further verification. We have treated the two sets as two separate events. In total, 56 accelerometers are used in the experiments. After filtering out invalid data (technical problems, participants leaving early, missing or incomplete questionnaires) we had valid data for 23 people in the first set and 21 in the second set. For our experiments, we used ~ 42 minutes from set I and ~ 22 minutes from set II.

Survey Responses. When a participant left the event, we asked them to fill in a questionnaire with the same six questions used for 'enjoyment' and 'immersion' at the 'Dance Performed' event. Questionnaires were taken separately for the two sets. Thus, a participant joining only one of these events filled in the relevant questionnaire only. For the first set, 48 percent of the participants stated they really enjoyed the drone presentation (delfly) while 62 percent of them chose the 'real' presentation of the surgical device as the top moment. For the second set, only 6 participants noted a favorite moment. These all consistently preferred *Zigeunerreisen*, the musical performance at the end of the presentation.

4 Data Analysis

In this section, we analyze the datasets in terms of *shared* experience and movement. Our assumption was that both the participants' subtle and more expansive movements are related to the experience of the event. In Section 5, we evaluate predicting *individual* questionnaire responses from measured movement.

We used the variance of the magnitude of the accelerometer readings, which are shown to act as the best proxy for the physical activity level of the participants in [46], using a sliding window of 2 seconds (40 samples) with 1 second shift (20 samples) to capture the subtle variations in motion while preserving a fine time scale. This window size is empirically found to perform best whereas larger window sizes suppresses subtle movements we are interested in. Before calculating the variance, the z-score of the magnitude is computed to remove interpersonal differences. Then, for each dataset, we computed the Mutual Information (MI) of this variance for every possible pair of participants, creating a pairwise co-occurrence measurement of the physical activity over time. These signals were computed over a sliding window with a size of 60 samples and shifted by one sample, resulting in a vector reflecting co-occurrence of motion, over time, between two participants.

4.1 Binary Labels for Evaluation

For our analysis, we convert the questionnaire responses for each experience category ("enjoyment", "immersion", "recommendation" and "mood improvement") to a binary label. We set up the questionnaires to contain three redundant questions per category and averaged the answers to obtain a single numerical value for each category. This was converted to a "positive" or "negative" experience: participants whose averaged answer was below 5 for a category were placed in the negative class for that category. This way, we obtain four different labels for each participant in Dataset 1, and two different labels in Dataset 2. The class distributions of all categories for each event obtained with this setup are given below.

Dataset 1. For "enjoyment" and "recommendation", the majority of participants (26 out of 32) gave positive answers. 22 participants

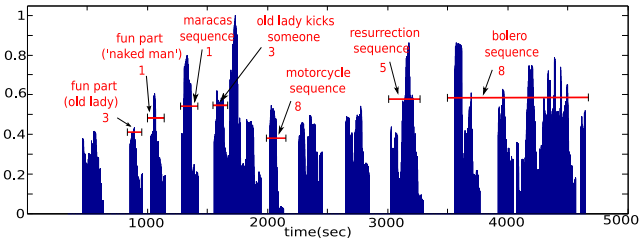

Fig. 1. Mean co-occurrence measurement distance over time for all participants using Mutual Information (MI) for Dataset 1. Moments that were reported as salient are highlighted in red, together with the number of times they were reported.

thought "the performance affected their mood positively". The distribution for the "immersion" task is relatively more balanced with 17 participants in the positive class.

Dataset 2. 21 out of 23 participants and 18 out of 20 participants gave positive responses to the "enjoyment" questions for the first and second sets. For "immersion", 16 out of 23 and 9 out of 20 participants responded positively for the first and second sets, respectively.

4.2 Dataset 1

We investigated three things: 1) Do moments when people move in synchrony correspond to salient moments of the performance? 2) Is the proximity between people in the audience a factor that also triggers synchronous motion and does it affect the reported experience? 3) Will it be possible to automatically identify sitting neighbors through proximity sensors?

4.2.1 Synchrony and Salient Moments

We hypothesized that salient moments should correspond to a high MI among all participants. We used an Otsu threshold [54] on the the mean pairwise MI of all possible pairs (computed as explained in Section 4) to select parts where co-occurrence of the physical activity is relatively high. Traditionally, Otsu thresholding is used for converting grayscale images (continuous pixel values from 0 to 1) to black and white (binary). Since our MI values also lay between 0 and 1, we employed this method to detect moments of high co-occurrence of physical activity. Fig. 1 shows a timeline depicting timesteps for which the average MI for all pairs is more than the threshold, in blue, as well as all reported favorite moments together with their reporting frequency, in red. Notice that all of the reported favorite moments show up in the MI, including the two moments declared as favorite for the majority of participants (*motorcycle* and *bolero finale*), and that most moments of high MI correspond to reported moments. This shows that memorable moments for people during these events can be captured by their coordinated movements, as they share the experience.

4.2.2 Impact of Proximity

In this section, we analyze the impact of proximity in the enjoyment of the event. The participants were seated throughout the performance, making people's relative location static. We identified where each participant was sitting during the performance and used this ground truth information for the analysis. Fig. 3 shows the mean MI (calculated over the whole event) between neighboring participants (side, front and back neighbors). In addition, red subjects represents those who did not enjoy the event while the green ones did.

Fig. 3 has 41 connections between neighbo-ring participants. Similar to the former analysis, MI between two people is considered low if the value is less than the Otsu threshold computed on all connected pairs. When all four neighbors are considered, there are 15 and 12 connections of high and low average MI between people who enjoyed the event, respectively. The values are 7 and 6 if only side neighbors are considered. Higher number of connections with high MI shows that proximity might have an effect on the evaluation but the low difference between numbers of high and low MI connections makes it harder to come up with hard conclusions.

We must also account for the people that came together to the event. The groups of participants that are known to come together to the event are shown in Fig. 3 as dashed black lines. Although the pairwise MI and enjoyment of the event is comparatively high for some of the participants that came together, this does not generalise for all groups of acquaintances. Also, there are five cases where two participants shared a high MI but their enjoyment of the event differed. We surmise that such high co-occurrence values are due to shared comments or other shared actions that had no relation with the performance, but we cannot directly prove this since we do not have video recordings of the audience.

4.2.3 Identifying Sitting Neighbors

In this section, we investigate whether we can leverage the proximity data to identify who is sitting close to whom. Basically, we are trying to see if it is possible to construct a connectivity graph similar to Fig. 3 automatically, using the proximity detections of our sensors. The proximity sensing is omnidirectional, however how the shielding effect of the body influences the detection of individuals sitting sideways, front or behind is unclear. Even assuming neighbors can be detected, it is unclear how far they can be sensed and how this relationship can be characterized since no signal-strength is recorded by the sensors.

One would assume that the closer two individuals sit together, within the detection range of the sensor of 2-3 meters, the more frequently their nodes will detect each other. With this assumption, we investigate which neighbors are frequently detected through sensing by the following methodology:

1) For every node $u_{i,j}$ (participant sitting at row i and column j), count how often each ID was detected over the duration of the event,
2) Keep top K IDs as the candidate neighbors,
3) Check if these K candidate neighbors correspond to:
   a) 1-Hop side neighbors ($u_{i,j-1}, u_{i,j+1}$)
   b) Front and back neighbors ($u_{i-1,j}, u_{i+1,j}$)
   c) 1 and 2-Hop side neighbors ($u_{i,j-1}, u_{i,j+1}, u_{i,j-2}, u_{i,j+2}$)
   d) Diagonal neighbors ($u_{i-1,j-1}, u_{i-1,j+1}, u_{i+1,j+1}, u_{i+1,j-1}$)

For evaluating cases a) and b), we set $K = 2$. For cases c) and d), $K = 4$. Frontal and diagonal neighbors yield low recalls of 0.37 and 0.24 respectively, while 1-hop neighbors yield precision of 0.62 and recall of 0.86. When we also add 2-hop neighbors, we obtain a precision of 0.59 and a recall of 0.84. These suggest that some of the neighbors detected for the 1-hop neighbors (with $K = 2$) are 2-hop neighbors (lowering the precision), but 2-hop neighbors are not consistently detected such that precision and recall are still similar with $K = 4$. The other source of error in precision in both cases are the rare detections of frontal and diagonal neighbors, which are not detected consistently but sometimes appear in the top-K list for some individuals.

To conclude, it is not possible to satisfactorily detect diagonal, front and back neighbors through proximity sensing. However, the precision and recall values obtained when classifying 1-hop and 2-hop neighbors show that it is possible to detect who is sitting at the sides of an individual with some sampling of frontal and diagonal neighbors. This information is valuable in analyzing events where people are seated but the seating arrangement is unknown.

4.3 Dataset 2

The same analysis of Section 4.2.1 was carried out for Dataset 2. Fig. 2 shows the mean MI among all participants along with the separations between the sections of the event (parts and talks). One key difference between the two datasets is their structure. Dataset 1 is collected in a continuously flowing event, whereas the Day of

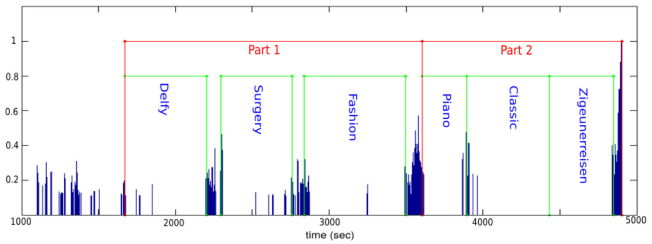

Fig. 2. Mean co-occurrence measurement distance over time for all participants using Mutual Information (MI) for Dataset 2. The two main sets of the event are highlighted in red and the talks in green.

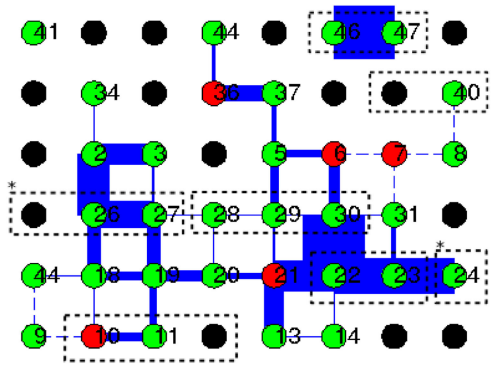

Fig. 3. Mean MI between participants sitting together during the Dataset 1. Green dots indicate subjects who did enjoy the performance, red dots indicate subjects who did not, and black dots indicate empty seats (or people for which no data is available). The width of the blue bars indicate the average MI value throughout the performance, while dashed lines are non relevant MI relations.

Wonder has clearly delimited talks. This structure can be clearly seen in 2, where after each talk a high MI value is observed, corresponding to the rounds of applause and possible relocations between talks. This behavior was not present in the Dataset 1 as that event only had a single round of applause at the end of the performance. We also see the highest peaks between the two talks and after the second talk ends. People were allowed to leave at these points, corresponding to global high co-occurrences of physical activity.

In contrast to Dataset 1, we don't see many peaks during the talks. Different factors can explain this. First of all, the crowd in this event was a mix of seated and standing people. This might cause an overall drop of the global pairwise MI, since the measured reactions of seated and standing people are expected to be different. Second and more importantly, there are many parts where everyone in the audience reacts, such as the ending of the talks. Such parts are shown to be have high global MI, and they might suppress co-occurring subtle responses to the event by increasing the threshold. So, if the aim is to find salient moments in an event like this, moments like applause or people leaving should be excluded from the analysis.

5 Automatic Prediction of the Evaluations

We investigate, on both datasets, whether it is possible to predict questionnaire responses about the performance from accelerometer data. In the following sections, we perform classification experiments, where we present our methodology for automatically predicting a participant's evaluation of the events.

5.1 Classifying Experience

5.1.1 Methodology

To emphasize the connection between the information contained in the motion data and the participants' experience of the event, in our classification experiments we focus on a simple set of features and a well-understood classifier. Our features are the acceleration variance along each axis and the overall acceleration magnitude variance. Our classifier is a Linear Support Vector Machine (SVM, [55]). Since the number of samples is limited, we opted for a model with few parameters. We evaluated the performance of our method with leave-one-participant-out cross validation. The hyperparameters of the SVM are selected using nested cross validation on the training set. The variance values of each window are treated as independent features, resulting in high dimensional feature vectors, but since we do not expect all intervals to be equally informative, we applied filtering to select the features from informative intervals. The steps of feature extraction, feature (interval) selection and classification are presented below:

Feature Extraction

1) For each participant, compute the variance of the acceleration X, Y, Z, and magnitude, using a 2s sliding window with 1s shift, resulting in 4 features for each 2s time window.
2) Concatenate the feature vectors from each window to obtain a single feature vector of the whole event, per participant.

Feature (Interval) Selection

1) Compute Dynamic Time Warping (DTW) values over the previously computed features for each pair of participants. The DTW window size is the number of feature extraction windows it contains, so that a 5-sample DTW window contains 20 features (4 features for each of the five 2s window).
2) Obtain an OTSU threshold using all computed DTW values.
3) Select the windows whose DTW scores exceed the threshold.
4) For each participant, keep the features of the selected windows. So, if 3 non-overlapping windows are selected with 5-sample DTW windows, each participant's resulting feature vector has 60 dimensions.

Classification

1) For further dimensionality reduction, apply Principal Component Analysis (PCA) to the feature vectors and keep the principal components which preserve 99 percent of the variance.
2) For each participant p:
   a) Train a Linear SVM on the feature vectors of all participants, excluding p.
   b) Classify the feature vector of p.

Our assumption is that the intervals with high average pairwise DTW distance are more discriminative than the rest. In an ideal scenario, intra-class distances should stay relatively stable throughout the event, so that intervals where the average DTW distance is high are those where the inter-class distances are maximized. We expect average DTW to provide better discrimination between classes than mutual information, as windows with high MI would correspond to moments where the classes would be almost indistinguishable and all participants' movements are synchronized. Empirical results using MI supported this claim, with performance scores significantly lower than the proposed method for the majority of the tasks.

Features (variance values) are computed over 4705, 2503 and 1293 windows for the Dataset 1 and Dataset 2 Parts 1 and 2, respectively. Each window corresponds to an interval of 2 seconds and 4 features. The number of remaining intervals after feature selection depends on the window size for the computation of the DTW values, where we experimented with window sizes ranging from 1 sample to 80 samples, each with a 1 sample shift. For Dataset 1, the number of selected intervals ranged from 44 to 1065. For the first and second parts of Dataset 2, number of selected intervals ranged from 166 to 802 and 55 to 935. After the PCA, dimensions of the feature vectors used in the classification experiments of Dataset 1 ranged between 18 and 28, whereas the range for Dataset 2 was 15 to 22.

TABLE 1
Prediction Performances for Both Datasets

Method\BAcc (%)	Enjoyment			Recomm.	Immersion			Mood
	D1	D2-1	D2-2	D1	D1	D2-1	D2-2	D1
DTW IS (1 Sample)	100^{**}	**48**	*50*	92^{*}	58	**58**	100^{**}	46
DTW IS (5 Sample)	100^{**}	**50**	*50*	100^{**}	65^{*}	**65**	100^{**}	47
DTW IS (10 Sample)	100^{**}	**48**	*50*	100^{**}	59	**68**	90^{**}	53
DTW IS (20 Sample)	100^{**}	**63**	*50*	92^{*}	65^{*}	**58**	90^{**}	56
DTW IS (40 Sample)	92^{**}	**53**	*47*	90^{*}	52	**71**	94^{**}	47
DTW IS (80 Sample)	81^{**}	**48**	*44*	73	52	**68**	84^{*}	49
Whole Event	48	**48**	*44*	65	46	**68**	*52*	51

($* \rightarrow p < 0.1$) ($** \rightarrow p < 0.05$).
*Scores for Dataset 2 parts 1 and 2 are shown in **bold** and* italic*, respectively, as second and third values at cells of "Enjoyment" and "Immersion".*

5.1.2 Results and Discussion

Table 1 reports the performance results for both datasets for different window sizes, both with and without pre-filtering salient intervals using thresholded DTW distance. We selected balanced accuracy [56] as our performance metric to account for the class imbalance. The results that are significantly better than using the whole event are indicated with an asterisk. Significance was computed using an asymptotic McNemar's test with misclassification costs that are inversely proportional to the class distributions. While training the Linear SVM, the samples are weighted inversely proportional to the class frequencies to combat imbalance.

Dataset 1. Without interval selection, the results (final row of Table 1) are generally unsatisfactory. Any task other than predicting "recommendation" has a balanced accuracy score at, or below, chance level. We should note that we did apply PCA to the feature vectors for the non-filtered method. Without interval selection, PCA requires many more components to keep the same amount of variance in order to model the many non-informative intervals, supporting our claim of interval selection is necessary.

We were able to get perfect classification results for "enjoyment" when performing interval selection, with window sizes ranging from 1 to 20 samples. In addition, all other window sizes still yielded significantly better performance ($p < 0.05$) than using the whole event or chance prediction. The performance tends to drop with increasing window size, suggesting a small window size might be more suitable for detecting enjoyment. Further supporting this claim, using data from the whole event fails to give results better than random. Even though computing DTW over single-sample windows might sound counter-intuitive, the filtering approach is still able to find informative intervals. This works probably because even a single sample has temporal information, since its value is extracted from a 2 second window.

Results for "recommendation" show similar characteristics to "enjoyment": perfect classification, significantly better than using the whole event ($p < 0.05$), is achieved with window sizes of 5 and 10 and the performance tends to drop with the increasing window size. Using features from the whole event still provides performance better than random with a balanced accuracy of 65 percent. This might simply mean that "recommendation" can be inferred from the whole event with an acceptable performance but some parts of the event might be still more indicative, providing finer results.

The performance for "immersion" and "mood" is relatively poor compared to the others. These experiences are less immediately about the performance itself, and may be harder to report objectively, bolstering the case for immediate sensing over reappraisal. For "immersion", the highest performance is 65 percent, obtained with 5 and 20 sample windows which is still significantly better than using the whole event ($p < 0.1$). The performance for this task does not seem to be changing too much between 1 to 20 samples, and fluctuates between 58 and 65. However, using larger windows result in poor performance. For "mood", the highest obtained performance is 56 percent with a window size selection of 20 samples. Most of the other window sizes resulted in performances worse than random.

The optimal window size tends to differ for each experience, suggesting that some experiences are reflected in shorter time scales than others. Also, most tasks performed best when small to medium sized windows were used, indicating that large window sizes fail to capture the connection between participants' movements.

We experimented with computing DTW distances on the raw accelerometer magnitude signal instead of the variance over a window. This experiment resulted in performance scores that were worse than random for "immersion" and "mood". Highest balanced accuracy scores for tasks of "enjoyment" and "recommendation" were 58 and 68 percent, respectively. For all tasks, using the variance rather than raw signal in DTW distance computation resulted in relatively better performance. We can conclude that variance in acceleration is a useful feature, both as a feature for prediction and for the interval selection using the thresholded DTW distance. This is probably because the variance of acceleration reflects the *amount* of movement rather than the precise movement and its direction, leading to more robust recognition.

Dataset 2. As shown in Table 1, for the first part of the event, we were able to obtain better-than-random performance for both tasks, but the very limited number of negative examples make it impossible to make hard conclusions. The highest performance for "enjoyment" was 63 percent, obtained with a window size of 20 samples. Compared to the balanced accuracy of 48 percent obtained with the whole event setup, this result supports pre-filtering with DTW. However, all other window sizes failed to capture any meaningful information, providing either slightly higher or lower performances than a random baseline. Compared to the results on Dataset 1, this suggests that the optimal window size for a task might also change with the characteristics of the event. For "immersion", the optimal window size seems to be quite arbitrary. The highest performance, 71 percent, is obtained with 40 samples. However, using features from the whole event also results in a balanced accuracy of 68 percent which is not significantly different than the best score. Thus, for the first part of this event, "immersion" can be detected with an acceptable performance without requiring filtering.

Results are quite different for the second part of the event. For "enjoyment", most of the window sizes resulted in a balanced accuracy of 50 percent, showing that the classifier fails to learn anything from the data. Multiple factors might have caused this. First, we only had 2 negative samples. We believe the negative samples for the first part were more informative than the second one, making it possible to obtain better performance. Second, the length of the second part is the shortest of our all datasets. In order to capture a complex concept such as enjoyment, temporally more extended data might be required. Finally, this part was the closing act. Even though the majority of people reported this part as one of their favorite, 1) there may be a memory effect in play, where people report the event that's most fresh in their mind as the favorite, and 2) movement patterns of people might tend to change when nearing the end of events.

We were able to get perfect classification for "immersion" with windows of 1 and 5 samples. Contrary to the first part, using the features from the whole event results in a balanced accuracy of 52 percent and the results with filtering are significantly better. This supports our claim that the optimal window size depends not only on the task, but also on external factors to the task.

These follow-up experiments with an event of differing characteristics show that whether people are standing or sitting does not really affect our capacity to analyze people's response to the event. Our proposed methodology still provides competitive results, even in the quite unruly, noisy, real-world situation of these festival-style events.

6 Conclusions

In our study, we have investigated how an audience's perception of a performance can be recognized and measured from their body movements with an accelerometer such as typically present in smart phones. We have presented our results on two datasets collected during live performances. These have different characteristics, both in terms of the performance itself and the audience demographics. Building on findings from appraisal theory and affective studies, that show how a stimulus creates an affective response which can be connected to experience, we analyzed whether subtle and complex concepts would be reflected in the body motion as measured by a simple accelerometer hung around the neck. These concepts included "enjoyment", "immersion", an improvement in mood as a result of the performance, and whether participants would "recommend" dance in general. Using the variance of the acceleration, we were able to predict the audience's self-reported experience in both events, in terms of the aforementioned complex concepts.

Importantly, joint coordination in the acceleration variance, which reflects how the body movements of participants are related, helps to distinguish salient from non-salient moments of the performance. Restricting the analysis to these leads to significant improvements over using each individual person's body movements from the entire performance period. We analyzed how the spatial layout of a seated audience might affect its members' experience of the performance and presented a proximity-based method that can automatically detect neighboring participants with satisfying performance. Our experiments shows huge promise in enabling us to measure the implicit responses of people while watching a live performance without the need for more traditional—and less practical—sensing approaches using physiological or brain signals. However, and perhaps more importantly, our experiments demonstrate the potential of quantifying the experience of 'a cultural night out', highlighting the relevance of the social context in moderating an individual's enjoyment of an event.

Acknowledgments

The authors would like to thank the Distributed & Interactive Systems group at CWI, the Lucent Theatre, Djana Eminovic, Flora Rajakowitsch, and Andrew Demetriou for their help and support in designing and executing the 'Dance Performance' and 'A day of Wonder' experiments. This publication was partially supported by the Dutch national program COMMIT, the European Commission under grant agreement number 601033 - MOnarCH, and the Costa Rican Institute of Technology.

References

[1] W. H. Whyte, *The Social Life Of Small Urban Spaces*. New York, NY, USA: Project for Public Spaces Inc, Jan. 1980.
[2] C. L. Nightingale, C. Rodriguez, and G. Carnaby, "The impact of music interventions on anxiety for adult cancer patients: A meta-analysis and systematic review," *Integrative Cancer Therapies*, vol. 12, no. 5, pp. 393–403, 2013.
[3] M. Ritter and K. G. Low, "Effects of dance/movement therapy: A meta-analysis," *Arts Psychotherapy*, vol. 23, no. 3, pp. 249–260, 1996.
[4] D. Fujiwara, L. Kudrna, and P. Dolan, "Quantifying and Valuing the Wellbeing Impacts of Culture and Sport," UK Department of Culture, Media and Sport, London, U.K., Apr. 2014, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/304899/Quantifying_and_valuing_the_wellbeing_impacts_of_sport_and_culture.pdf
[5] T. Nenonen, R. Kaikkonen, J. Murto, and M.-L. Luoma, "Cultural services and activities: The association with self-rated health and quality of life," *Arts Health*, vol. 6, no. 3, pp. 235–253, 2014.
[6] A. C. Michalos and P. M. Kahlke, "Arts and the perceived quality of life in british columbia," *Social Indicators Res.*, vol. 96, no. 1, pp. 1–39, 2010.
[7] K. R. Scherer, A. Schorr, and T. Johnstone, *Appraisal Processes in Emotion: Theory, Methods, Research*. London, U.K.: Oxford Univ. Press, 2001.
[8] M. Soleymani and M. Pantic, "Human-centered implicit tagging: Overview and perspectives," in *Proc. IEEE Int. Conf. Syst. Man Cybern.*, 2012, pp. 3304–3309.
[9] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1424–1445, Dec. 2000.
[10] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[11] M. Karg, A.-A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić, "Body movements for affective expression: A survey of automatic recognition and generation," *IEEE Trans. Affect. Comput.*, vol. 4, no. 4, pp. 341–359, Oct.-Dec. 2013.
[12] A. Kleinsmith and N. Bianchi-Berthouze , "Affective body expression perception and recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 15–33, Jan.-Mar. 2013.
[13] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2007, pp. 71–82.
[14] H. Gunes and M. Piccardi, "Automatic temporal segment detection and affect recognition from face and body display," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)*, vol. 39, no. 1, pp. 64–84, Feb. 2009.
[15] M. Karg, K. Kuhnlenz, and M. Buss, "Recognition of affect based on gait patterns," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)*, vol. 40, no. 4, pp. 1050–1061, Aug. 2010.
[16] S. D'Mello and A. Graesser, "Automatic detection of learner's affect from gross body language," *Appl. Artif. Intell.*, vol. 23, no. 2, pp. 123–150, 2009.
[17] C. Chênes, G. Chanel, M. Soleymani, and T. Pun, "Highlight detection in movie scenes through inter-users, physiological linkage," in *Social Media Retrieval*. Berlin, Germany: Springer, 2013, pp. 217–237.
[18] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan.-Mar. 2012.
[19] J. Fleureau, P. Guillotel, and I. Orlac, "Affective benchmarking of movies based on the physiological responses of a real audience," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 73–78.
[20] I. Cohen, N. Sebe, A. Garg, M. S. Lew, and T. S. Huang, "Facial expression recognition from video sequences," in *Proc. IEEE Int. Conf. Multimedia Expo*, vol. 2, pp. 121–124, 2002.
[21] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 46–53.
[22] Y.-L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," in *Handbook of Face Recognition*. Berlin, Germany: Springer, 2005, pp. 247–275.
[23] N. Sebe, M. S. Lew, Y. Sun, I. Cohen, T. Gevers, and T. S. Huang, "Authentic facial expression analysis," *Image Vis. Comput.*, vol. 25, no. 12, pp. 1856–1863, 2007.
[24] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proc. 13th Annu. ACM Int. Conf. Multimedia*, 2005, pp. 677–682.
[25] H. G. Wallbott, "Bodily expression of emotion," *Eur. J. Social Psychology*, vol. 28, no. 6, pp. 879–896, 1998.
[26] J. K. Burgoon, L. K. Guerrero, and K. Floyd, *Nonverbal Communication*. Evanston, IL, USA: Routledge, 2016.
[27] D. Bernhardt and P. Robinson, "Detecting affect from non-stylised body motions," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2007, pp. 59–70.
[28] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2007, pp. 48–58.
[29] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic recognition of non-acted affective postures," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)*, vol. 41, no. 4, pp. 1027–1038, Aug. 2011.
[30] A. S. Brown and J. L. Novak, *Assessing the Intrinsic Impacts of a Live Performance*. San Francisco, CA, USA: WolfBrown, 2007.
[31] M. Reason and D. Reynolds, "Kinesthesia, empathy, and related pleasures: An inquiry into audience experiences of watching dance," *Dance Res. J.*, vol. 42, no. 2, pp. 49–75, 2010.
[32] L. Bennett, "Patterns of listening through social media: Online fan engagement with the live music experience," *Social Semiotics*, vol. 22, no. 5, pp. 545–557, 2012.
[33] M. R. Lockstone, L. Olga Junek, S. Hudson, and R. Hudson, "Engaging with consumers using social media: A case study of music festivals," *Int. J. Event Festival Manag.*, vol. 4, no. 3, pp. 206–223, 2013.
[34] A. Leask, A. Hassanien, and P. C. Rothschild, "Social media use in sports and entertainment venues," *Int. J. Event Festival Manag.*, vol. 2, no. 2, pp. 139–150, 2011.
[35] "USA Today. Providence theater experiments with 'tweet seats'," 2013. [Online]. Available: http://www.usatoday.com/story/tech/2013/01/27/theater-tweet-seats/1868693/
[36] B. Bläsing, B. Calvo-Merino, E. S. Cross, C. Jola, J. Honisch, and C. J. Stevens, "Neurocognitive control in dance perception and performance," *Acta Psychologica*, vol. 139, no. 2, pp. 300–308, 2012.
[37] E. S. Cross, L. Kirsch, L. F. Ticini, and S. Schütz-Bosbach, "The impact of aesthetic evaluation and physical ability on dance perception," *Frontiers Human Neuroscience*, vol. 5, 2011, Art. no. 102.
[38] B. Calvo-Merino, C. Jola, D. E. Glaser, and P. Haggard, "Towards a sensorimotor aesthetics of performing art," *Consciousness Cognition*, vol. 17, no. 3, pp. 911–922, 2008.

[39] C. J. Stevens, H. Winskel, C. Howell, L.-M. Vidal, J. Milne-Home, and C. Latimer, "Direct and indirect methods for measuring audience reactions to contemporary dance," *Dance Dialogues: Conversations Across Cultures, Artforms and Practices: Proc. World Dance Alliance Global Summit*, Brisbane, pp. 13–18, Jul. 2008, https://ausdance.org.au/articles/details/direct-and-indirect-methods-for-measuring-audience-reactions-to-contem
[40] C. Latulipe, E. A. Carroll, and D. Lottridge, "Love, hate, arousal and engagement: Exploring audience responses to performing arts," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2011, pp. 1845–1854.
[41] C. Wang, E. N. Geelhoed, P. P. Stenton, and P. Cesar, "Sensing a live audience," in *Proc. 32nd Annu. ACM Conf. Human Factors Comput. Syst.*, 2014, pp. 1909–1912.
[42] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SigKDD Explorations Newslett.*, vol. 12, no. 2, pp. 74–82, 2011.
[43] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Proc. Int. Conf. Pervasive Comput.*, 2004, pp. 1–17.
[44] C. Doukas, I. Maglogiannis, P. Tragas, D. Liapis, and G. Yovanof, "Patient fall detection using support vector machines," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innovations*, 2007, pp. 147–156.
[45] T. Zhang, J. Wang, P. Liu, and J. Hou, "Fall detection by embedding an accelerometer in cellphone and using KFD algorithm," *Int. J. Comput. Sci. Netw. Security*, vol. 6, pp. 277–284, 2006.
[46] G. Englebienne and H. Hung, "Mining for motivation: Using a single wearable accelerometer to detect people's interests," in *Proc. 2nd ACM Int. Workshop Interactive Multimedia Mobile Portable Devices*, 2012, pp. 23–26.
[47] X. Bao, S. Fan, A. Varshavsky, K. Li, and R. Roy Choudhury, "Your reactions suggest you liked the movie: Automatic content rating via reaction sensing," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, pp. 197–206.
[48] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J. Pinton, and A. Vespignani, "Dynamics of Person-to-Person interactions from distributed RFID sensor networks," *Plos One*, vol. 5, no. 7, Jul. 2010, Art. no. e11596. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0011596
[49] C. Martella, A. van Halteren, M. van Steen, C. Conrado, and J. Li, "Crowd textures as proximity graphs," *IEEE Commun. Mag.*, vol. 52, no. 1, pp. 114–121, Jan. 2014.
[50] D. Roggen, M. Wirz, D. Helbing, and G. Tröster, "Recognition of crowd behavior from mobile sensors with pattern analysis and graph clustering methods," *Netw. Heterogeneous Media*, vol. 6, pp. 521–544, 2011.
[51] E. W. See-To, S. Papagiannidis, and V. Cho, "User experience on mobile video appreciation: How to engross users and to enhance their enjoyment in watching mobile video clips," *Technological Forecasting Social Change*, vol. 79, no. 8, pp. 1484–1494, 2012.
[52] T. Schubert, F. Friedmann, and H. Regenbrecht, "The experience of presence: Factor analytic insights," *Presence: Teleoperators Virtual Environments*, vol. 10, no. 3, pp. 266–281, 2001.
[53] H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 1, pp. 50–69, 2010.
[54] R. Gonzalez and R. Woods, *Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice Hall, 2008.
[55] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011, [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[56] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *Proc. 20th Int. Conf. Pattern Recognit.*, 2010, pp. 3121–3124.